Implementasi Algoritma *Suffix Tree Clustering* dan *Nearesrt Neighbor* untuk Mengelompokkan Berita pada Timeline Twitter

Implementation of Algoritma Suffix Tree Clustering and Nearesrt Neighbor to Classify News in the Twittier Timeline

Jumadi*[1], Edi Winarko[2]
UIN Sunan Gunung Djati Bandung ; Jl. A.H. Nasution 105 Cipadung Bandung
Universitas Gadjah Mada; Skip Utara Jl. Kaliurang Yogyakarta

Jurusan Teknik Informatika, UIN Sunan Gunung Djati, Yogyakarta
Jurusan Ilmu Komputer dan Elektronika, FMIPA UGM, Yogyakarta e-mail:
*[1]jumadi@if.uinsgd.ac.id , [2]ewinarko@ugm.ac.id

Abstrak

Kedinamisan konten *tweet* berita yang disebarkan oleh organisasi penyedia berita pada *Twitter*, menimbulkan banyaknya jumlah *tweet* yang dipublikasikan setiap harinya. Hal ini dapat menambah panjang halaman web mikroblog, sehingga menimbulkan permasalahan klasik yaitu memerlukan proses *page scrolling* yang lebih lama pada saat proses pembacaan semua teks *tweet* berita yang ada. Salah satu pemecahan permasalahan yang dapat dilakukan untuk mengurangi panjang halaman web tersebut adalah dengan cara melakukan pengelompokkan teks berita yang ada secara tematik. Sistem pengelompokan yang tepat untuk permasalahan ini adalah sistem pengklasteran. Berdasarkan beberapa penelitian yang ada, salah satu metode yang baik dalam proses pengklasteran dokumen teks adalah *Suffix tree Tree Clustering* (STC). Metode ini mempunyai tingkat ketepatan yang sangat tinggi karena dalam pembentukan klaster berdasarkan pada *phrase-shared* di antara dokumen-dokumen teks yang ada.

Tetapi salah satu penelitian yang ada, dalam melakukan proses pengklasteran dengan menggunakan algoritma *Suffix Tree Clustering (STC)* masih menghasilkan dokumen teks anggota klaster *Other Topics* dalam jumlah yang banyak dan jika diperhatikan dokumen teks anggota klaster ini juga diketahui masih ada relevansinya dengan dokumen teks anggota pada klaster yang ada. Oleh karena itu, dokumen teks yang berada pada klaster *Other Topics* ini, perlu dibandingkan dengan semua dokumen teks di klaster yang ada untuk mengetahui tingkat kemiripannya. Dengan demikian, dokumen teks anggota klaster *Other Topics* ini dapat diklasifikasikan ke dalam salah satu klaster tertentu dengan menggunakan fungsi *cosine similarity* berdasarkan dari hasil perhitungan dengan menggunakan metode *Vector Space Model* (VSM) yang mengacu pada frekuensi term dan frekuensi dokumen yang ada. Hasil perhitungan ini, akan digunakan oleh metode *Nearest Neighbor* dalam proses klasifikasi untuk menentukan klaster tujuan perpindahan bagi dokumen teks anggota klaster *Other Topics*. Kriteria klaster sebagai tujuan perpindahan adalah klaster dengan jumlah anggota terbanyak yang memiliki kemiripan tertinggi. Proses perpindahan dokumen teks anggota klaster *Other Topics* ini akan mengakibatkan berkurangnya jumlah angota klaster tersebut. Pada akhirnya, jika klaster *Other Topics* tidak memiliki anggota maka klaster ini pun dapat dihilangkan.

Kata kunci: Classification, Clustering, Cosine Similarity, Nearest Neighbor, Suffix Tree Clustering.

Abstract

Dynamism news tweet content are disseminated by news organization providers on Twitter, causing the large number of tweets published every day. It can increase the length of Microblog web pages, and inflict to the classic problems that require page scrolling process is longer during the process of reading all of the existing text news tweets. The problems solving that can be done to reduce the length of the web pages is by grouping the existing text news thematic. Grouping system suitable for this problem is the clustering system. Based on some existing research, one good method in the process of clustering text documents is a Suffix Tree Clustering (STC). This method has a very high accuracy rate, because clusters create based on phrase-shared among documents existing text.

But one of the existing research in the process of clustering algorithms using STC, still produce text documents Other Topics cluster members in large numbers and text documents members of this cluster are still relevant to the text document members of the existing clusters. Therefore, the text documents that is in the Other Topics cluster need to compare with all text documents in the existing clusters to determine the level of similarity. Thus, a text document Other Topics cluster members can be classified into one particular cluster by using the cosine similarity function based on the results of calculations using the method of Vector Space Model (VSM) which refers to the term frequency and the frequency of existing documents. Results of this calculation will be used by the Nearest Neighbor method in the classification process to determine the destination cluster displacement for text documents Other Topics cluster members. The main criteria of goal cluster as destination of displacement is the cluster with the highest number of members that have the highest similarity. The process of moving text document cluster members Other Topics impact on the reduction in the number of members of this cluster. Finally, if the Other Topics cluster has no members then this cluster can be eliminated.

Keyword: Classification, Clustering, Cosine Similarity, Nearest Neighbor, Suffix Tree Clustering.

1. Pendahuluan

Berdasarkan penelitian yang dilakukan oleh Zamir dan Etzioni (1998), algoritma yang digunakan untuk melakukan pengklastran dokumen web kali pertama adalah *suffix tree clustering (STC)*, algoritma klasterisasi ini memiliki waktu linear dalam mengelompokkan dokumen hasil pencarian ke dalam group-group berdasarkan kata atau frase yang terdapat di dalam dokumen yang ada. Kemudian Osiński dan Weiss (2004), mengembangkan *open source framework* dengan nama *Carrot*^{2}. Kesuksesan dan popularitas aplikasi *Carrot2* adalah mengorganisir hasil dari pencaraian di internet agar lebih mudah dalam menjelajah dalam bentuk pengelompokkan secara tematik hasil pencarian pada saat menggunakan browser internet, yang dikenal dengan proses klasterisasi dan STC adalah salah satu algoritma yang digunakan dalam proses pengklasteran.

Tetapi, kinerja algoritma klasterisasi STC yang dikembangkan oleh *Carrot*^{2} masih memiliki kekurangan, yaitu sering dijumpai hasil pengklasteran pada dokumen anggota klaster *other topics* dalam jumlah banyak dibandingkan dengan klaster yang ada dan jika diperhatikan kata-kata yang membentuk dokumen teks anggota klaster o*ther topics*, terdapat kemiripan dengan kata-kata teks pada klaster-klaster yang ada. Sehingga memungkinkan dokumen teks anggota klaster o*ther topics* untuk dipindahkan ke salah satu dari klaster-klaster yang ada berdasarkan kemiripan.

Mengacu pada konsep yang dibahas oleh Liao (2002), untuk mengatasi permasalahan ini perlu adanya proses klasifikasi dokumen teks *twitter* yang berada di klaster *other topic*

dengan cara menghitung nilai kemiripan antar dokumen dengan fungsi *cosine similarity* berdasarkan frekuensi term dan frekuensi dokumen yang ada, sesuai dengan konsep algoritma *vector space model*. Hasil dari perhitungan ini, kemudian digunakan oleh metode *nearest neighbor* untuk menentukan klaster dengan jumlah anggota terbanyak yang memiliki kemiripan sebagai klaster tujuan proses klasifikasi. Dengan demikian anggota klaster *other topics* akan berkurang bahkan habis sehingga klaster ini pun dapat dihilangkan.

2. Metode Penelitian

Metode yang digunakan pada penelitian ini meliputi:

1. Objek penelitian
   Pengelompokkan teks secara tematik pada status *tweet* atau *retweet* berita pada *twitter* yang didapat dari akun tertentu atau hasil pencarian dengan kata kunci tertentu, menggunakan algoritma s*uffix tree clustering* (STC) untuk proses klasterisasi, sedangkan pengelompokkan anggota klaster *other topics* hasil dari algoritma STC agar terklasifikasi ke klaster yang telah ada lainnya, dengan menggunakan algoritma klasifikasi *Nearest Neighbor.*
2. Data yang diperlukan
   a. Data primer
      Praproses mendapatkan inputan data berupa teks yang berasal dari teks *tweet* dan *retweet* penggalan berita pada status di *Twitter*. Proses pengambilan data ini menggunakan pustaka *LinqtoTwitter*. Data status ini berisi konten teks, waktu penebitan teks, *screen name* dan *image profile* pengguna.
   b. Data sekunder
      Algoritma *Suffix Tree Clustering* memproses masukan teks *tweet* dari *Twitter* dan menghasilkan nama-nama klaster beserta nama-nama dokumen yang menjadi anggotanya. Salah satu klaster berlabel *Other Topics*, dengan metode *Nearest Neighbor* anggota klaster ini akan diubah ke klaster yang ada.
3. Teknik pengumpulan data
   a. Observasi
      Penelitian ini menitikberatkan pada proses pengubahan nama klaster anggota klaster *Other Topics* yang dihasilkan oleh algoritma *Suffix Tree Clustering* menggunakan metode *Nearest Neighbor.* Dokumen teks berita ini berasal dari teks *tweet* pada *Twitter.*
   b. Studi Pustaka
      Mempelajari hasil penelitian-penelitian lain yang telah ada dan penelitian tersebut melibatkan algoritma *Suffix Tree Clustering* dan *Nearest Neighbor* dalam pengelompokan dokumen teks.
   c. Metode pengembangan sistem
      1) Pengambilan dokumen teks *tweet* atau *retweet* menggunakan pustaka *LinqtoTwitter* untuk mendapatkan teks, *image profile url*, dan waktu serta *user screen name*.
      2) Pra-proses meliputi penghapusan *stopword* dan *stoplist*, proses *tokenizing* dan *stemming* dengan menggunakan *Porter stemming for Bahasa Indonesia*
      3) Pembentukan klaster menggunakan pustaka *Carrot*^{2} dengan algortima *Suffix Tree Clustering* (STC).
      4) Pengklasifikasian anggota klaster *Other Topics* hasil dari proses algoritma STC, menggunakan metode *Nearest Neighbor* (NN)

5) Proses perhitungan pajang vektor untuk fungsi *vector space model* (SVM) menggunakan fungsi TF/IDF
6) Proses perhitungan kemiripan antar dokumen teks, menggunakan fungsi *cosine similarity*
7) Visualisasi hasil pengklasteran dan klasifikasi, ditampilkan dalam aplikasi bebasis web dan konsul (*console*)

3. Hasil dan Pembahasan

3.1 Data hasil praproses

Teks *tweet* berita yang dipublikasikan oleh para jurnalis Indonesia di akun *twitter* milik perusahaannya, dijadikan sebagai data yang akan diolah pada penelitian ini. Tahap awal dalam pengolahan data teks *tweet* berita adalah penghapusan *stopword*, *stoplist* dan proses *stemming* dalam Bahasa Indonesia. Hasil praproses terhadap 50 teks *tweet* berita yang didapat dari akun *twitter* @kompascom pada tanggal 26 September 2013, 5 dari 50 teks yang ada dapat dilihat pada Tabel 1.

Tabel 1. Hasil praposes

No	Teks tweet	Hasil proposes
1	IHSG Menunggu Momen Pembalikan Arah http://t.co/eByre0ngaJ	ihsg tunggu momen balik arah
2	Pink Star, Berlian Termahal di Dunia http://t.co/YPg05OsSgV	pink star berlian mahal dunia
3	Pencabutan Pentil Meluas, Kontainer Juga Kena http://t.co/PW6W6jCuzW	cabut pentil luas kontainer kena
4	Pulau Baru Muncul di Pakistan Setelah Gempa http://t.co/UT1xDBWfnb	pulau muncul pakistan gempa
5	Jalan Berbayar di Jakarta Bisa Dipercepat http://t.co/KBZvWeNeN6	jalan bayar jakarta bisa cepat

a. Data berasal dari twitter berdasarkan nama akun

Hasil proses pengklasteran dengan menggunakan algoritma *suffix tree clustering* dan melibatkan praproses pada Bahasa Indonesia berdasakan data yang diambil dari twitter dengan nama akun @kompascom, menghasilkan 6 klaster termasuk klaster *"other topics"* dengan jumlah 38 buah anggota.

b. Data berasal dari twitter berdasarkan kata kunci

Proses pengklasteran dilakukan juga pada data teks twitter yang diambil dengan cara memasukan kata kunci "seminar internsional" dapat dilihat pada Gambar 1. Jumlah klaster yang dihasilkan sebanyak 9 klaster termasuk klaster *"other topics"* dengan jumlah 18 anggota. Pada proses selanjutnya, yaitu proses klasifikasi. Semua anggota klaster *"other topics"* sebanyak 18 buah ini, akan diklasifikasikan ke dalam 8 klaster yang ada berdasarkan kemiripan teks.

Gambar 1. Hasil pengklasteran teks twitter berdasarkan kata kunci

3.2 Hasil klasifikasi anggota klaster *Other Topics*

Pada Gambar 2 menunjukkan proses klasifikasi, sedangkan Gambar 3 merupakan hasil dari proses klasifikasi dengan menggunaan metode *nearest neighbor* terhadap anggota teks klaster *other topics* hasil dari proses pengklasteran dengan menggunakan algortima *suffix tree clustering*. Nilai *k* yang digunakan pada metode *nearest neighbor* ini adalah 4, lebih jelasnya dapat dilihat pada Gambar 2.

Gambar 2 Proses klasifikasi *nearest neighbor* dengan *k* bernilai 4

Proses klasifikasi yang ditunjukkan pada Gambar 2 dapat dilihat tabel yang berisi informasi kode klaster, kode dokumen dan nilai kemiripan. Kode klaster C2 menunjukkan klaster nomor 2 memiliki dokumen dengan nomor 17, 22 dan 31 mempunyai kemiripan dengan dokumen nomor 1 sebesar 0,006589 dan klaster nomor 1 memiliki dokumen dengan nomor 38 dengan nilai kemiripan sebesar 0,001556. Dengan demikian klaster dengan kode C2 memiliki 3 dokumen yaitu D17, D22, D31 yang memiliki kemiripan dengan dokumen D1 sebagai anggota klaster *"other topics"* dan klaster berkode C1 memiliki sebuah dokumen dengan kode D38 yang memiliki kemiripan dengan dokumen D1. Ada 4 dokumen dokumen yang dilihat nilai kemiripan, menunjukan bahwa *k* pada algoritma nearest neighbor bernilai 4. Ekspresi ini, lebih dikenal dengan notasi k=4.

Berdasrkan penjelasan yang ada, dapat diketahui bahwa klaster C2 memiliki jumlah dokumen sebanyak 3 dokumen sedangkan klaster C1 terdapat 1 dokumen yang memiliki kemiripan dengan dokumen D1 sebagai anggota klaster "*other topics*". Dengan demikian, klaster C2 dengan nama "*Dlm, Bahas Dlm Seminar Converence*" merupakan klaster tujuan perpindahan dokumen D1 dalam proses klasifikasi. Hasil proses klasifikasi semua anggota klaster "*other topics*" dapat dilihat pada Gambar 3.

Gambar 3 Hasil proses klasifikasi

Hasil proses klasifikasi teks anggota klaster "*other topics"* dengan menggunakan metode *nearest neighbor* dari 50 dokumen teks twitter yang didapat dengan memasukan kata kunci "seminar internasional" dalam pencarian teks tweet. Jumlah klaster yang dihasilkan dari proses klasterisasi adalah 9 klaster termasuk klaster *other topics* dengan jumlah anggota sebanyak 18 dokumen. Jumlah anggota klaster *other topics* setelah melalui proses klasifikasi menjadi 0. Pada Tabel 2 dapat diketahui jumlah anggota klaster setelah mengalami proses klasifikasi.

Tabel 2 Hasil klasifikasi anggota klaster *other topics*

No	Nama Klaster	Jumlah Anggota	
		A	B
1	Seminar Implementasi, Emas 2045, Sumbar	8	9
2	Dlm, Bahas Dlm Seminar Converence	9	24
3	Smentara Kuota Sdh Penuh	4	4
4	IAIN Seminar	10	10
5	2045 Whit Ohio State	3	3
6	Terang Dlm Bijaksa Panitia	3	3
7	Ngikut Seminar Tamrin Benerrrrr	3	3
8	Panitia	5	7
9	Other Topics	18	0

Keterangan :
A adalah proses pengklasteran dengan algoritma s*uffix tree clustering* B adalah proses klasifikasi dengan algoritma *nearest neighbor*

3.3 Pengujian hasil klasifikasi

Proses pengujian bertujuan untuk mengecek hasil proses klasifikasi menggunakan metode *nearest neighbor* dengan hasil proses klasifikasi yang dilakukan secara manual terhadap semua dokumen teks pada klaser *other topics* yang berpindah ke klaster lain yang ada. Proses klasifikasi dengan metode *nearest neighbor* dilakukan oleh sistem dengan cara memilih klaster tertentu yang memiliki banyak anggota yang mempunyai kemiripan dengan teks anggota klaster *other topics*. Sedangkan proses klasifikasi yang dilakukan secara manual, dilakukan dengan cara mengecek kesesuaian semua konten anggota klaster *other topics* dengan label klaster-klaster yang ada. Hasil proses klasifikasi teks anggota klaster *other topics* yang dilakukan oleh sistem maupun yang dilakukan secara manual dapat dilihat pada Tabel 3.

Tabel 3 Rekap perubahan kelompok anggota klaster *other topics*

No (1)	A (2)	B (3)	C (4)	D (5)	E (6)	F (7)
1	D1	C2	D17, D22, D31	0.006	C8, C7, C2	-
2	D3	C2	D17, D22, D31	0.006	C4, C8, C2	√
3	D4	C2	D17, D22, D31	0.006	C8, C7	-
4	D7	C2	D17, D22, D31	0.005	C7, C2	√
5	D13	C1	D44, D41	0.214	C7, C8	-
6	D14	C2	D17, D22, D31	0.009	C4, C8	-
7	D15	C2	D17, D22, D31	0.016	C8, C7	-
8	D19	C2	D17, D22, D31	0.016	C3, C8	-
9	D32	C2	D17, D22, D31	0.007	C1,C8, C2	√
10	D33	C2	D17, D22, D31	0.007	C1, C8, C2	√
11	D35	C2	D17, D22, D31	0.006	C8, C7, C2	√
12	D36	C2	D17, D22, D31	0.016	C3, C8	-
13	D37	C2	D17, D22, D31	0.117	C6, C8, C2	√
14	D39	C8	D34	0.001	C8,C6	√
15	D40	C8	D34	0.001	C6, C8	√
16	D47	C2	D17, D22, D31	0.009	C8, C6, C2	√
17	D48	C2	D17, D22, D31	0.005	C1, C8	-
18	D49	C2	D17, D22, D31	0.006	C1, C8, C2	√

Keterangan:
A adalah kode dokumen klaster *other topics*
B adalah kode klaster tujuan perpindahan dokumen
C adalah kode dokumen klaster tujuan yang memiliki kemiripan
D adalah nilai *similarity* dokumen *other topics* terhadap dokumen pada klaster tujuan
E adalah klaster rekomendasi tujuan perpindahan secara manual
F adalah hasil pengecekan kesesuain B dan E (kolom 2 & 5)

Berdasarkan data pada Tabel 3 jumlah teks anggota klaster *other topics* ada 11 dari 18 dokumen yang memiliki kesesuain antara konten teks anggota klaster *other topics* dengan label nama klaster yang dijadikan tujuan perpindahan klaster. Jadi persentasi ketepan hasil klasifikasi *nearest neighbor* adalah 61,11%.

4. Kesimpulan

Berdasarkan penelitian yang telah dilakukan dapat disimpulkan beberapa hal sebagai berikut:

1. Hasil pengujian hasil proses klasifikasi yang dilakukan oleh sistem dengan menggunakan algoritma *nearest neighbor* dan hasil klasifikasi yang dilakukan secara manual dengan cara mengecek kesesuaian konten dokumen *other topics* terhadap makna yang terkandung pada label atau nama klaster tujuan perpindahan terhadap hasil klasifikasi adalah 61,11%.
2. Keakuratan pengelompokkan dokumen teks Bahasa Indonesia sangat dipengaruhi oleh kelengkapan daftar *stopword* dan *stoplist* serta proses *stemming* untuk teks *twitter* Bahasa Indonesia.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Bapak Edi Winarko. dan semua pihak yang telah mendukung kegitan penelitian ini. Semoga Alloh S.W.T. membalas dengan berlipat ganda kebaikan.

Daftar Pustaka

Arifin, A. Z., Darwanto, R., Navastara, D. A., Ciptaningtyas, H. T., 2008, *Klasifikasi Online Dokumen Berita dengan Menggunakan Algoritma Suffix Tree Clustering*, Seminar Sistem Informasi Indonesia, ITS, Surabaya.

Esko, U., 1995, *On-Line Construction of Suffix Trees*. In: Algorithmica, Vol. 14, No. 3., pp. 249-260.

Farach. M., 1997,*Optimal Suffix Tree Construction with Large Alphabets*, In Proc. 38th Annual Symposium on Foundations of Computer Science , pages 137–143. IEEE

Liao Y., 2002, *Review of K-Nearest Neighbor Text Categorization Method*, https://www.usenix.org/legacy/events/sec02/full_papers/liao/liao_html/node4.html, diakses 21 Agustus 2013

Pughazendi, N. dan M., Punithavalli, 2011, Temporal Databases and Frequent Pattern Mining Techniques, *International Journal of P2P Network Trends and Technology*, July to Aug Issue 2001, pp. 13 - 17

Weiner, P., 1973, *Linear pattern matching algorithms*, in Proceedings of the 14th Annual IEEE Symposium on Switching and Automata Theory, pp. 1–11,

Weiss D., Osinki S., 2004, *Carrot^2 Clustering Framework*, Poznan University of Technology, Poznan Poland

Wicaksono T., 2012, *Text Mining untuk Pencarian Dokumen Bahasa Inggris menggunakan Suffix Tree Clustering*, Jurusan Teknik Informatika, ITS

Winarko, E. dan Roddick, J.F., 2005, Discovering Richer Temporal Association Rule from Interval-Based Data: Extended Report, *Data Warehouse and Knowledge Discovery, LNCS*, 3589, pp. 315 – 325

Yusuf, Y.W., Pratikto, F.R., dan Gerry, T., 2006, Penerapan Data Mining dalam Penentuan Aturan Asosiasi Antar Jenis Item, *Proceding SNATI*, pp. E-53 – E-56, 1907-5022.

Zamir, O., Etzioni, O., 1998, *Web document clustering: a feasibility demonstration*. In: SIGIR 1998, pp. 46-54.